HALLUCINATION EVALUATION

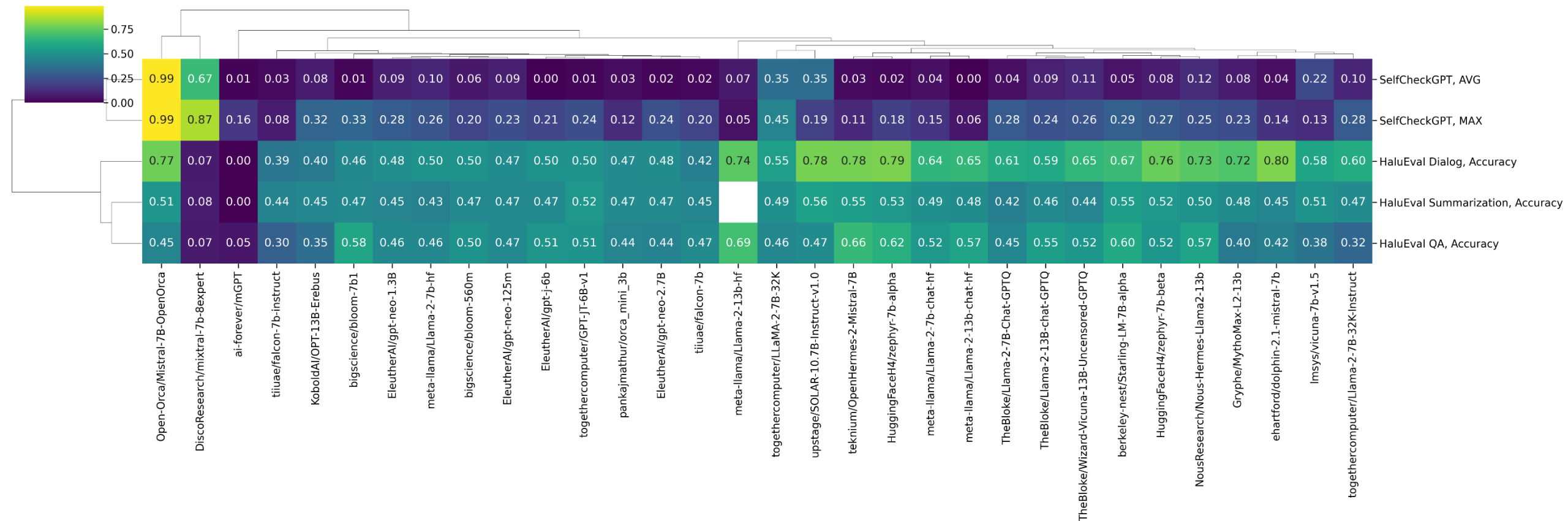## IN AI SYSTEMS

HOLOSCAN

Have you had any issues with the authenticity of the AI's responses?

Unraveling truthfulness in AI

Metric	Open-Orca/Mistral-7B-OpenOrca	DiscoResearch/mixtral-7b-8expert	ai-forever/mGPT	tiiuae/falcon-7b-instruct	KoboldAI/OPT-13B-Erebus	bigscience/bloom-7b1	EleutherAI/gpt-neo-1.3B	meta-llama/Llama-2-7b-hf	bigscience/bloom-560m	EleutherAI/gpt-neo-125m	EleutherAI/gpt-j-6b	togethercomputer/GPT-JT-6B-v1	pankajmathur/orca_mini_3b	EleutherAI/gpt-neo-2.7B	tiiuae/falcon-7b	meta-llama/Llama-2-13b-hf	togethercomputer/LLaMA-2-7B-32K	upstage/SOLAR-10.7B-Instruct-v1.0	teknium/OpenHermes-2-Mistral-7B	HuggingFaceH4/zephyr-7b-alpha	meta-llama/Llama-2-7b-chat-hf	meta-llama/Llama-2-13b-chat-hf	TheBloke/Llama-2-7B-Chat-GPTQ	TheBloke/Llama-2-13B-chat-GPTQ	TheBloke/Wizard-Vicuna-13B-Uncensored-GPTQ	berkeley-nest/Starling-LM-7B-alpha	HuggingFaceH4/zephyr-7b-beta	NousResearch/Nous-Hermes-Llama2-13b	Gryphe/MythoMax-L2-13b	ehartford/dolphin-2.1-mistral-7b	lmsys/vicuna-7b-v1.5	togethercomputer/Llama-2-7B-32K-Instruct
SelfCheckGPT, AVG	0.99	0.67	0.01	0.03	0.08	0.01	0.09	0.10	0.06	0.09	0.00	0.01	0.03	0.02	0.02	0.07	0.35	0.35	0.03	0.02	0.04	0.00	0.04	0.09	0.11	0.05	0.08	0.12	0.08	0.04	0.22	0.10
SelfCheckGPT, MAX	0.99	0.87	0.16	0.08	0.32	0.33	0.28	0.26	0.20	0.23	0.21	0.24	0.12	0.24	0.20	0.05	0.45	0.19	0.11	0.18	0.15	0.06	0.28	0.24	0.26	0.29	0.27	0.25	0.23	0.14	0.13	0.28
HaluEval Dialog, Accuracy	0.77	0.07	0.00	0.39	0.40	0.46	0.48	0.50	0.50	0.47	0.50	0.50	0.47	0.48	0.42	0.74	0.55	0.78	0.78	0.79	0.64	0.65	0.61	0.59	0.65	0.67	0.76	0.73	0.72	0.80	0.58	0.60
HaluEval Summarization, Accuracy	0.51	0.08	0.00	0.44	0.45	0.47	0.45	0.43	0.47	0.47	0.47	0.52	0.47	0.47	0.45		0.49	0.56	0.55	0.53	0.49	0.48	0.42	0.46	0.44	0.55	0.52	0.50	0.48	0.45	0.51	0.47
HaluEval QA, Accuracy	0.45	0.07	0.05	0.30	0.35	0.58	0.46	0.46	0.50	0.47	0.51	0.51	0.44	0.44	0.47	0.69	0.46	0.47	0.66	0.62	0.52	0.57	0.45	0.55	0.52	0.60	0.52	0.57	0.40	0.42	0.38	0.32

The complexity of evaluating and optimizing AI today

Highly experimental and manual inputs

- Automatic translation
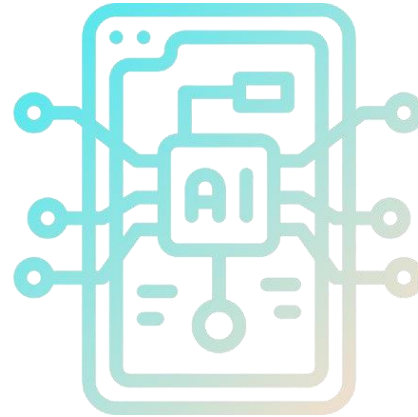- Prompt templates
- Processing chains
- Specific training data
- Hyperparameter settings
- Fine-tuning techniques
- Evaluation methods
- And more

Outputs evaluated essentially by human judgment
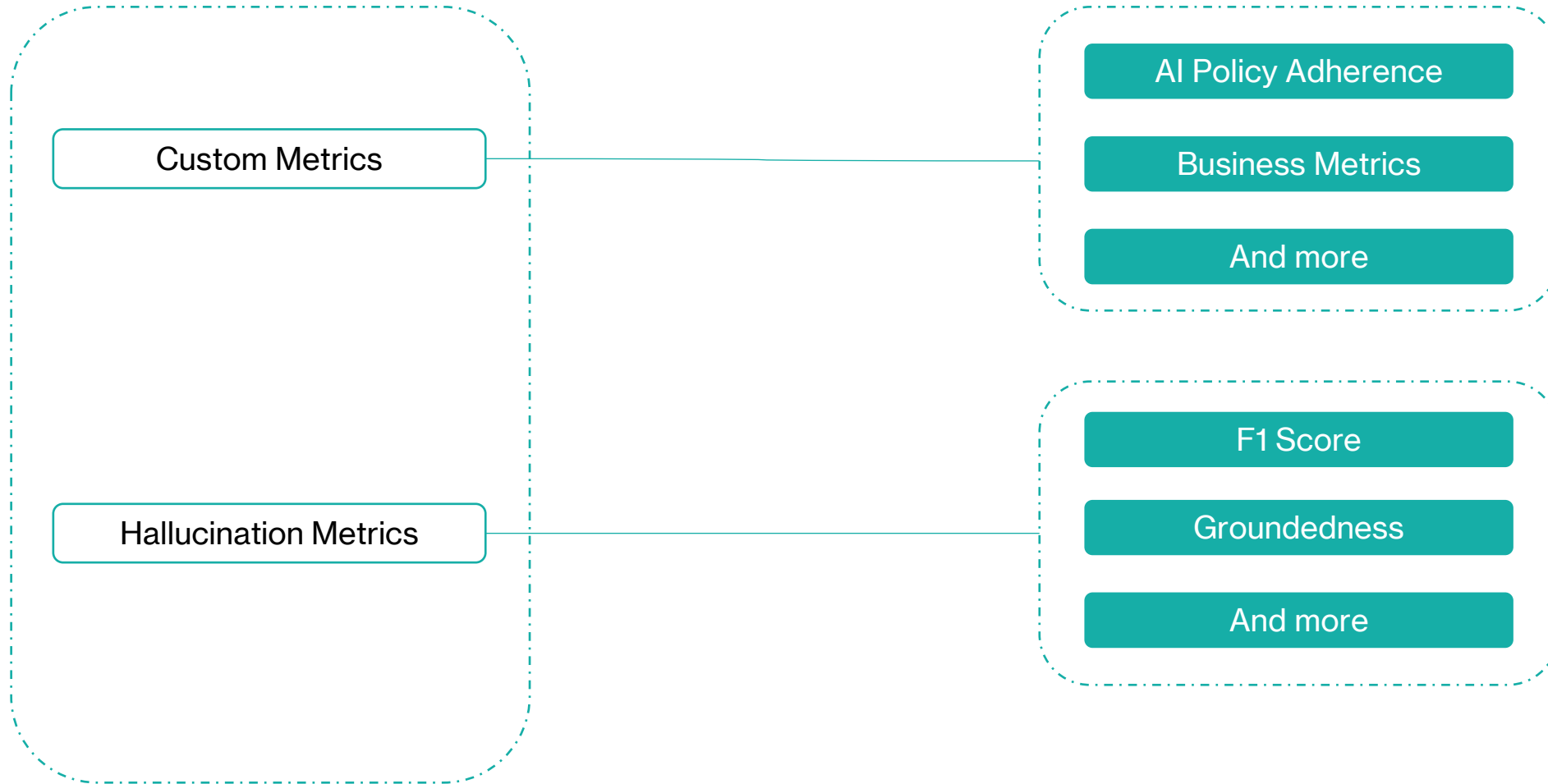
- Chat
- QnA
- Sentiment analysis
- Text generation
- Code generation
- Automatic translation
- Text summarization
- And more

Output evaluation metrics

Custom Metrics

- AI Policy Adherence
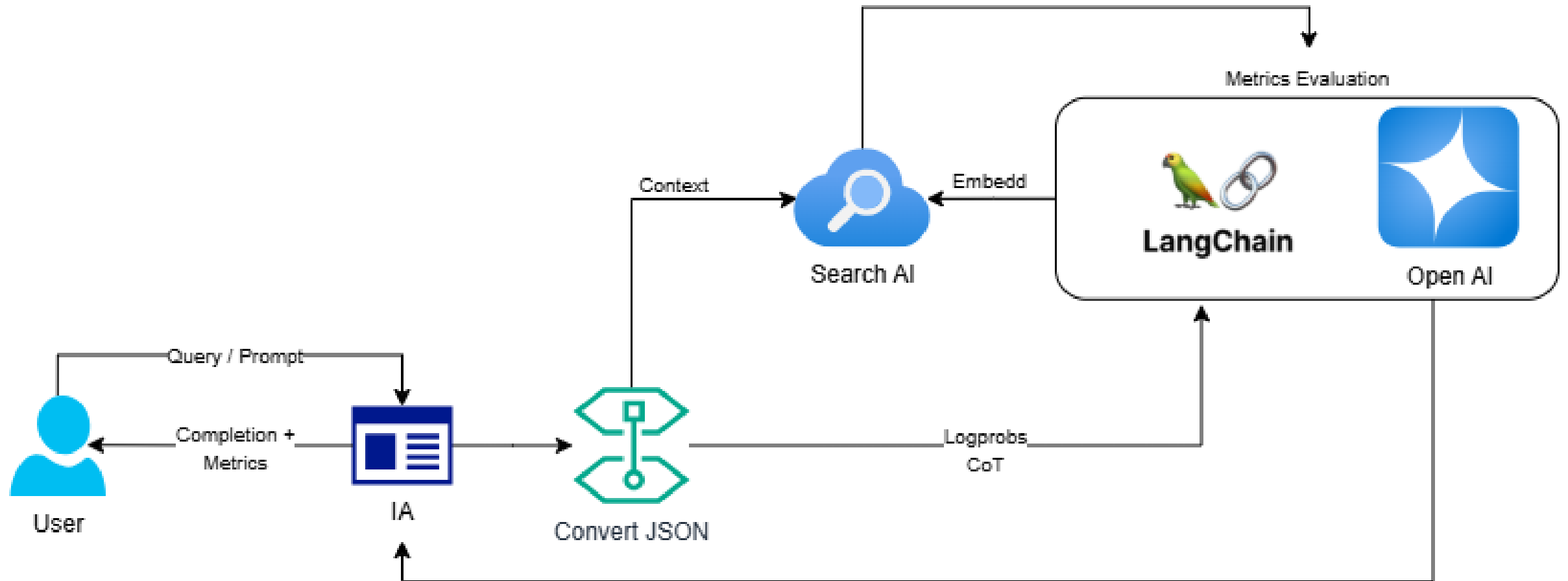- Business Metrics
- And more

Hallucination Metrics

- F1 Score
- Groundedness
- And more

AI evaluation metrics that are...

🎯 More precise

☑ Scalable to a variety of real-world tasks

⧗ Accelerate prompt engineering

💰 Economical

⚡ Minimal latency

Architecture

Demo

HoloScan´s Team

Carlos Nieto

BI Analyst
QA Tester

Luis Cerelli

Product Manager
Data Scientist

Adrian Panduro

AI Engineer
Data Engineer

Thanks!